

Обучение машинного перевода без параллельных текстов

Артеменков А.А., Гончаров М.Ю. Ярошенко А.М., Иванов А.В.,
Мазуров М.Ю., Борисова А.В., Скиднов Е.Д., Строганов Ф.А.

December 10, 2018

Список литературы



Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato.
Unsupervised machine translation using monolingual corpora only.
arXiv preprint arXiv:1711.00043, 2017.



Alex Graves and Jurgen Schmidhuber.
Framewise phoneme classification with bidirectional lstm and other
neural network architectures.
Neural Networks, 18(5-6):602–610, 2005.



Yunsu Kim and Jiahui Geng Hermann Ney.
Improving unsupervised word-by-word translation with language
model and denoising autoencoder.
2018.



Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th annual meeting on association for
computational linguistics*, pages 311–318. Association for
Computational Linguistics, 2002.

Постановка задачи

Рассматривается задача построения модели перевода текста без использования параллельных текстов, т.е. пар одинаковых предложений на разных языках.

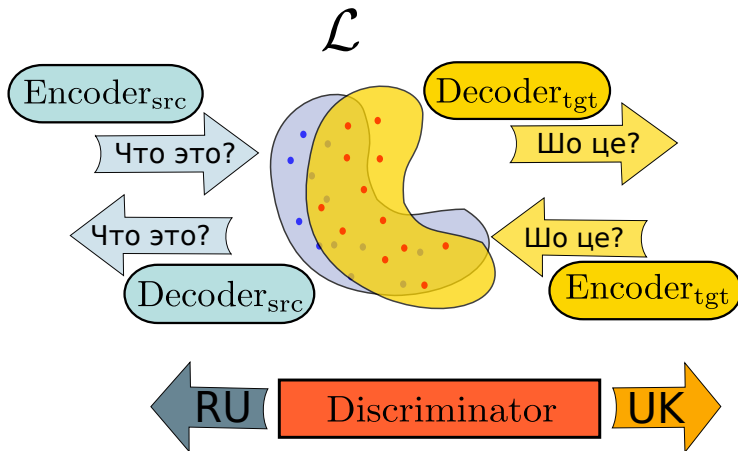
Пример.

- Что это?
- Вы говорите на украинском языке?

- Що це?
- Ви говорите українською мовою?

Данная задача возникает при построении моделей перевода для низкоресурсных языков (т.е. языков, для которых данных в открытом доступе немного).

Описание алгоритма



Описание эксперимента

Для украинско-русской пары:

- Непараллельная выборка: мультязычные статьи из Wikipedia.
- Параллельный корпус: OpenSubtitles-2018.

Для французско-английской пары:

- Параллельный корпус: multi30k.
- Метрика измерения качества: BLEU.

Обучение проводилось на обеих выборках, валидация на параллельном корпусе.

Результаты экспериментов

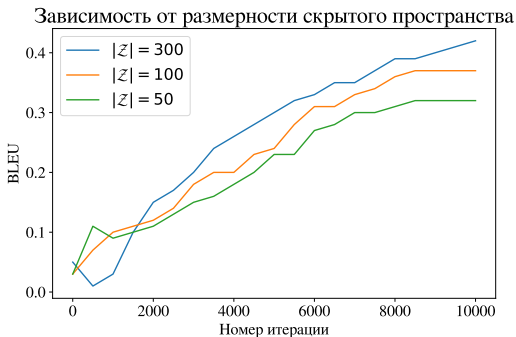


Рисунок: Зависимость BLEU от размерности скрытого пространства.

Результаты экспериментов

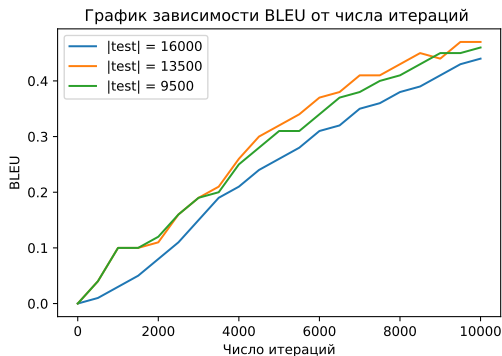


Рисунок: Зависимость BLEU от размера обучающей выборки.

Перевод с русского на украинский язык и обратно

Проблема: в языках слова могут иметь очень много форм. Из-за этого процесс обучения крайне долгий, вид кривых обучения не меняется кардинально при изменении модели.

decoder_size=500,attn_size=20,discr_size=20.png

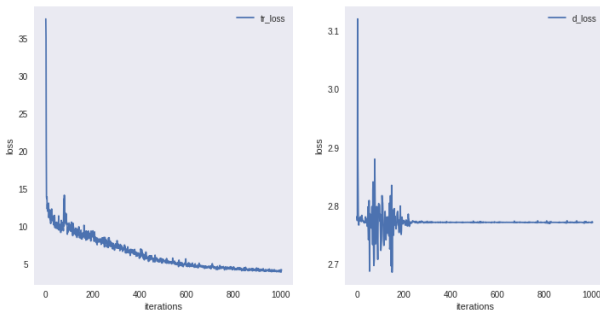


Рисунок: Вид кривых обучения на выборке русский-украинский

Применение стемминга

Стемминг — процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова.

WIKI, STEMMING: decoder_size=800,attn_size=70,discr_size=300

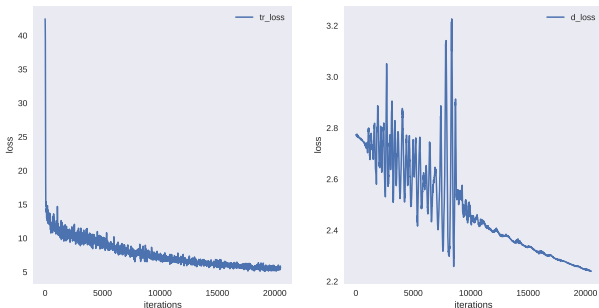


Рисунок: Кривые обучения в случае, когда всем словам были удалены последние две буквы

Результаты эксперимента

Table: Результаты, достигнутые для русско-украинского перевода.

Модель(decoder,discriminator,attention)	BLEU после 10^5 итераций
Простая: 100,100,50	0.113
Промежуточная: 200,300,50	0.219
Промежуточная: 800,600,150	0.222
Сложная: 2000,2000,150	<u>0.293</u>
Промежуточная: 800,300,70 + <i>STM</i>	0.413
Сложная: 2000,2000,150 + <i>STM</i>	0.427

- Качество найденных русско-украинских выборок не является достаточным как для построения хорошей модели, так и для валидации
- Была найдена одна из возможных проблем, связанная со структурой языков

Дальнейшие планы:

- Решение проблемы большого числа словоформ и слов не из словаря
- Поиск выборки с переводами более высокого качества